Exploration textométrique du corpus des dossiers de Bouvard et Pécuchet

Alexei Lavrentiev, Serge Heiden

***Revue Flaubert*, n° 13, 2013 | « Les dossiers documentaires de *Bouvard et Pécuchet* » : l'édition numérique du creuset flaubertien.**
Actes du colloque de Lyon, 7-9 mars 2012
Numéro dirigé par Stéphanie Dord-Crouslé

Exploration textométrique du corpus des dossiers de *Bouvard et Pécuchet*

Alexei Lavrentiev
CNRS – UMR ICAR

Serge Heiden
ENS de Lyon – UMR ICAR

Voir [Résumé]

La communication que nous avons présentée au colloque « *Bouvard et Pécuchet* : les "seconds volumes" possibles » témoigne d'une expérience d'exploitation du corpus des dossiers de *Bouvard et Pécuchet*[1] (dorénavant abrégé en « corpus Bouvard » ou « corpus DBP ») dans un cadre méthodologique et technologique très éloigné du projet d'origine. Elle a montré les possibilités offertes par l'usage d'un système d'encodage de sources textuelles ouvert et interdisciplinaire tel que proposé par XML et les recommandations du consortium de la Text Encoding Initiative (TEI), mais aussi les limites que posent l'extrême variabilité des pratiques d'encodage en TEI et la difficulté de concilier la représentation documentaire très précise de la source primaire avec la constitution de structures sémantiques pertinentes pour l'analyse textométrique.

Méthode textométrique et plateforme TXM

La textométrie est une méthode d'exploration de sources textuelles développée dans la lignée de la lexicométrie des années 1980. Ce qui se limitait avec la lexicométrie à l'exploitation du décompte de mots et à l'analyse de quelques contrastes entre groupes de textes a évolué avec la textométrie vers une exploitation basée sur toutes les informations disponibles pour chaque mot, comme son lemme ou sa description morphosyntaxique (substantif, verbe[2]...), toujours en tenant compte de son contexte d'apparition et de ses propriétés mais en y adjoignant les possibilités de contrastes internes entre plans textuels (entre sections, niveaux de discours, distinction corps du texte/apparat critique[3] ...). La méthode combine des outils quantitatifs de contraste (comme le calcul de la liste des mots les plus spécifiques à un sous-corpus donné), de cartographie d'ensemble (comme la représentation factorielle d'un ensemble de textes d'après la fréquence des mots qu'ils contiennent) ou de recherche de cooccurrents (mots les plus attirés par un mot donné dans un contexte défini par une fenêtre de mots ou une structure textuelle) avec des outils qualitatifs offrant des services de recherche plein texte de motifs de séquences de mots dont les résultats sont présentés sous forme de concordances ou de lecture hypertextuelle des éditions de textes du corpus.

Il existe un corpus conséquent d'applications de ces méthodes aux études littéraires[4]. Cependant cet article ne fera qu'effleurer les possibilités d'application dans ce domaine en se centrant surtout sur la relation entre le corpus traité et l'outil d'analyse TXM en vue d'études potentielles.

La plateforme TXM[5] est une implémentation ouverte de la textométrie[6] initiée lors du projet ANR Textométrie en 2007-2010. Elle fédère les différents algorithmes de la méthode auparavant disponibles dans divers outils (Hyperbase, Lexico, Alceste, Weblex...) dans une plateforme mutualisée dont les sources sont disponibles en ligne librement. De la sorte, les calculs peuvent être vérifiés de façon complètement transparente et améliorés de façon communautaire[7]. Elle est utilisable sous la forme d'une application pour poste (sous Windows, Mac OS X ou Linux) qui importe les sources des corpus pour ensuite proposer une interface d'outils d'analyse, ainsi que sous la forme d'un portail web qui offre la possibilité de mettre les corpus en ligne tout en fournissant une interface d'outils d'analyse par le biais d'un navigateur web connecté au portail.

Ce qui rend la plateforme TXM pertinente pour le corpus DBP est sa capacité à importer et analyser des sources textuelles encodées en XML-TEI. La section suivante discute en détail la façon d'importer ce corpus dans l'outil pour pouvoir l'exploiter.

Préparation et importation du corpus DBP dans la plateforme TXM

Le travail de préparation de corpus consiste à interpréter la structure de données du corpus telle qu'établie par le projet scientifique en des termes correspondant au modèle de données de l'analyse textométrique et à mettre en place une procédure d'importation qui permette de l'exploiter dans les meilleurs conditions possibles. Les outils de transformation de documents XML que l'on peut appliquer lors de l'importation dans TXM permettent de restructurer très profondément les données d'origine ; il est cependant souhaitable de limiter autant que possible la modification de la structure du corpus afin de préserver une certaine cohérence avec le projet initial. Dans les paragraphes qui suivent, nous commencerons par fournir un aperçu général de la structure des données du corpus DBP, avant de proposer son interprétation en termes de modèle de données de TXM. Enfin, nous décrirons les difficultés que nous avons rencontrées pour établir la correspondance et comment elles ont été surmontées.

Aperçu global

Le corpus fourni par le projet Bouvard se présente sous la forme de 3473 fichiers XML-TEI[8], dont chacun correspond à une page physique de l'archive numérisée. Certaines métadonnées les concernant se trouvent directement dans les fichiers (dans l'entête TEI et dans les attributs de la balise <text>), d'autres sont enregistrées dans une base de données externe. Certaines pages n'ont pas encore été transcrites et sont représentées par des « pages fantômes » : des fichiers XML-TEI avec des métadonnées, mais sans corps du texte. D'autres pages sont issues directement de la reconnaissance optique des caractères imprimés au sein des images de pages (OCR) et contiennent de nombreuses erreurs.

Le balisage du corpus est très riche et complexe : dans l'ensemble du corpus, on trouve 297 471 balises TEI de 64 types différents. Certains types de balises sont très fréquents (des dizaines de milliers d'occurrences), d'autres n'apparaissent que très rarement (moins d'une dizaine d'occurrences). Sur le plan sémantique, les balises servent à identifier des blocs de texte (, <ab>, <p>, <lg>, <l>, etc.), à marquer des segments ayant une mise en forme particulière (<seg>, <hi>, etc.), à proposer des corrections éditoriales ou des résolutions d'abréviations (<corr>, <reg> et <expan> regroupés avec <sic>, <orig> et <abbr> dans <choice>), à indiquer des termes à indexer (<term>), etc. Le balisage privilégie clairement l'objectif documentaire : représenter aussi précisément que possible la matérialité du document et assurer le lien avec le facsimilé numérique. Cet état hétérogène du balisage et de la qualité de numérisation du corpus représente une bonne opportunité pour tester la solidité des outils de préparation de corpus et la pertinence de certains outils génériques de TXM.

Quant au modèle de données de TXM, celui-ci est relativement simple : un corpus se compose d'un certain nombre d'« unités textuelles » (au moins une) qui sont composées à leur tour de chaînes d'« unités lexicales » (mots ou ponctuations). Des structures intermédiaires peuvent s'intercaler entre ces deux types d'unités (des sections, des segments de texte de toutes sortes balisés systématiquement ou ponctuellement).

Unité textuelle

Les unités textuelles sont normalement renseignées et qualifiées par un

ensemble de métadonnées (par exemple, le titre, l'auteur, la date ou toute autre caractéristique typologique du texte) qui permettent d'effectuer des analyses contrastives sur un corpus construit pour une étude donnée. Dans le cas d'un corpus d'œuvres littéraires publiées, il est relativement simple d'identifier les unités textuelles, même si des questions se posent pour des recueils de poésie ou pour des romans très volumineux. Dans le cas de DBP, cette question est beaucoup plus délicate et mérite une analyse attentive.

En effet, il ne s'agit pas vraiment d'un ensemble de textes mais de dossiers préparatoires où des notes de lecture prises par Flaubert se mélangent avec des citations manuscrites ou littéralement coupées-collées à partir d'autres ouvrages, ou encore avec des éléments de scénarios sans parler d'annotations postérieures de la main de Flaubert ou de son secrétaire. Certaines notes s'étendent sur plusieurs pages et forment ce qu'on peut appeler des « textes », mais la reconstitution de ces textes n'est pas acquise dans le corpus que nous traitons, car ce corpus est censé donner au chercheur qui l'utilise le pouvoir de composer librement différents parcours de lecture et d'interprétation.

Il ne peut donc s'agir à ce stade que d'une approche purement documentaire et pragmatique. La stratégie la plus simple serait de respecter la granularité d'origine, c'est à dire de considérer une page (un fichier de transcription) comme une unité textuelle. Ce choix est d'autant plus légitime que la principale métadonnée typologique encodée dans le corpus, à savoir le type de page (« scénarique », « note », « document » ou « synthèse ») porte sur les pages.

Cependant, si une note continue d'une page à l'autre, on risque de ne plus retrouver les contextes entourant les mots qui se trouvent à la fin de la première et au début de la seconde page, ce qui gênerait leur lecture dans une concordance par exemple. Par ailleurs, dans son état actuel, la gestion de corpus de plusieurs milliers d'unités textuelles par TXM (même toutes petites) n'est pas optimale : l'importation est très longue et certaines commandes peuvent provoquer des erreurs. Il était donc souhaitable d'envisager le regroupement des pages dans des unités plus grandes, ce que nous avons fait. Les « dossiers thématiques » dans lesquels les pages des dossiers documentaires du roman sont physiquement archivées et qui constituent la base de leur classement « patrimonial » se prêtent naturellement à ce rôle d'unité de regroupement. La réorganisation a pu être effectuée à partir du classement patrimonial sous forme de document XML et à partir des documents source grâce à une feuille de transformation XSLT 2. L'opération a généré 52 documents, dont un (le manuscrit C du *Dictionnaire des idées reçues*) n'a pas de contenu textuel, car il n'a pas encore été du tout transcrit. Les pages sont devenues des éléments XML <page> héritant des attributs « type » et « subtype », ce qui permet de les transformer en « structures intermédiaires » lors de l'importation dans TXM, et d'exploiter par la suite leur classement typologique.

Unité lexicale

L'unité lexicale se trouve à la base de toute analyse textométrique ou presque. Les dimensions du corpus et de ses parties sont calculées en nombre d'unités lexicales, ces unités portent des annotations telles que le lemme et la catégorie morphosyntaxique (qui peuvent être effectuées automatiquement lors de l'importation du corpus) ; de nombreuses mesures statistiques s'appuient sur ces unités et leurs séquences. Il est possible d'importer dans TXM un corpus où les unités lexicales sont déjà pré-balisées et étiquetées, mais le plus souvent la tâche de la segmentation lexicale est confiée à un outil automatique intégré dans le processus d'importation.

Deux types de problèmes peuvent se poser lors de la reconnaissance des unités lexicales par cet outil au moment de l'importation d'un corpus contenant des balises XML : des erreurs de segmentation (lorsqu'un mot se trouve coupé en plusieurs unités) et des erreurs d'enchaînement (lorsque des variantes d'un mot sont considérées comme deux mots distincts dans la chaîne textuelle). En ce qui concerne les erreurs de segmentation, elles peuvent survenir en cas de coupure de mot en fin de ligne ou de page, en cas de mise en relief ou de correction d'une partie d'un mot (par exemple, une grande lettrine, une lettre en exposant ou une lettre omise ajoutée par le transcripteur). Pour éviter les erreurs de segmentation lors de l'importation dans TXM, il convient de pré-baliser les mots qui posent problème en utilisant la balise TEI <w>. Il est également possible d'effectuer un tel pré-balisage « à la volée » lors de la procédure d'importation.

Le cas des sauts de ligne « à l'intérieur » des mots peut être relativement facile à traiter à condition que ceux-ci soient marqués d'une manière régulière.

La TEI propose depuis peu un mécanisme standard à cette fin (l'attribut « break » de la balise <lb>), mais cet attribut n'était pas disponible au moment de la définition de schéma de balisage du projet Bouvard. La solution retenue a été de baliser le caractère de césure (le trait d'union). Dans le cas où le document source ne présente aucun caractère de césure, la reconstitution des mots coupés devient difficile. Par ailleurs, il arrive que dans le document source les traits d'union soient dupliqués ou placés uniquement au début de la nouvelle ligne, ce dont les transcriptions du projet Bouvard tiennent compte. Enfin, sur certaines pages, les traits d'union n'étaient pas balisés au moment où nous avons obtenu le corpus, ce qui complexifie encore le traitement automatisé. Néanmoins, en traitant les cas repérables grâce à des expressions régulières et des conditions XPath[9], on peut éliminer une part importante des erreurs.

Plus rarement, on trouve à l'intérieur des mots des balises de corrections éditoriales (<supplied> pour des lettres omises dans la source et restituées) ou de mise en forme (<hi> pour des lettres en exposant ou parties de mots soulignées). Ces cas peuvent aussi être traités automatiquement à condition qu'ils soient repérables sur des critères formels. Par exemple, on peut s'appuyer sur l'usage de l'espace blanc devant et après ces balises, ce qui comporte tout de même un certain risque, car les éditeurs et les outils de traitement XML peuvent parfois indenter les documents en insérant automatiquement des sauts de ligne et des tabulations devant les balises pour faciliter la lecture ce qui change la disposition des espaces.

En ce qui concerne l'enchaînement textuel, la TEI propose depuis la version P5 un mécanisme puissant pour encoder côte à côte des formes erronées et des corrections, des formes originales et des régularisations, des abréviations et leurs résolutions en plaçant les balises correspondantes dans un élément <choice>. Le modèle de données de cet élément est très souple : on peut encoder plus de deux alternatives et même imbriquer des <choice>, et il n'y a pas de contrainte quant à la granularité linguistique de ce qui est balisé : une partie d'un mot, un mot ou plusieurs mots. Tous ces cas de figure se trouvent effectivement dans le corpus DBP. Évidemment, une telle souplesse pose de sérieux problèmes pour la reconstitution de la chaîne textuelle. La solution la plus simple est de choisir l'une des alternatives et d'éliminer les autres (par exemple, ne retenir que les formes corrigées, régularisées et désabrégées), mais des informations précieuses pour l'analyse peuvent ainsi être perdues.

Nous avons tenté de préserver les « formes alternatives » en les transférant dans le modèle TXM sous forme de propriétés d'unités lexicales (par exemple, la forme abrégée est codée dans la propriété « txm-abbr » de la forme complète). Notre traitement repose sur l'hypothèse que le contenu des différentes balises regroupées est équivalent à une unité lexicale, ce qui est le cas dans la grande majorité des occurrences. Les erreurs générées par les autres cas peuvent être détectées au cours de l'analyse du corpus par TXM et corrigées au coup par coup dans les sources.

Comme nous l'avons déjà indiqué, les unités lexicales peuvent porter des annotations sous forme de « propriétés ». La catégorie morphosyntaxique et le lemme sont ajoutés automatiquement lors de l'importation du corpus si un outil de TAL correspondant (en l'occurrence TreeTagger) est activé dans TXM et s'il dispose d'un modèle linguistique adéquat. Pour le corpus DBP, nous avons utilisé le modèle linguistique du français moderne mis à disposition avec le TreeTagger. Faute d'avoir le temps d'évaluer précisément les performances de TreeTagger sur le corpus DBP, nous avons vérifié les 100 associations « forme – catégorie – lemme » les plus fréquentes pour les noms et les adjectifs et nous n'avons détecté que très peu d'erreurs apparentes en dehors de caractères spéciaux (faute de balisage particulier, les astérisques ont été traités comme des unités lexicales et étiquetés « nom » ou « adjectif ») : le pronom indéfini *un* a été étiqueté « nom » et la préposition latine *in* a été étiquetée « adjectif ».

En plus des annotations automatiques, il est possible d'équiper les unités lexicales de propriétés supplémentaires qui peuvent être utiles dans l'exploitation du corpus. Dans le cas du corpus DBP, il semble que l'information sur la « main » responsable de la rédaction d'un segment de texte soit particulièrement importante. Un attribut spécial « dbp:hand » a été introduit dans le schéma d'encodage du projet Bouvard. Cet attribut peut être placé à tout niveau : une page entière, un fragment, une note, une correction ou un caractère. Afin de pouvoir contraster facilement les usages linguistiques des différents scripteurs, il nous a paru utile de projeter cette information sur chaque mot du corpus dans sa propriété lexicale « dbp-hand ».

Structures intermédiaires et plans textuels

Le niveau des structures intermédiaires est facultatif dans le modèle de données TXM, mais il peut être très utile, voire indispensable dans l'analyse de corpus richement balisés, comme c'est le cas du corpus DBP. Le rôle de ces structures peut être double : découper l'unité textuelle pour permettre des analyses contrastives plus fines que sur les textes entiers et séparer ce qu'on appelle les « plans textuels ».

En ce qui concerne le découpage des textes, la structure intermédiaire principale du corpus DBP est la page physique. Puisque la page est l'unité principale du corpus d'origine, il est certain que chaque unité lexicale se trouve sur une page ou une autre. Les pages sont systématiquement annotées en type et sous-type, ce qui permet de créer des contrastes intéressants lors de l'exploitation du corpus.

Au niveau inférieur à la page, la structuration du corpus DBP est plus complexe : 168 pages contiennent des divisions et sous-divisions de types variés (« article », « lettre », « texte » ou « print »), 1384 pages sont composées directement de listes (balise <list>), dont 6 listes bibliographiques (<listBibl>) et 1650 pages sont divisées en blocs de texte (balise <ab>). Ces structures ne peuvent donc pas être utilisées pour partitionner le corpus dans son ensemble, mais il est possible de créer des sous-corpus correspondants.

À un niveau encore inférieur, on trouve des segments de texte (quelques mots ou phrases) balisés pour différentes raisons. Mis à part les balises présentant des variantes d'encodage d'un segment donné regroupées sous <choice>, que nous traitons au niveau des unités lexicales, il s'agit de balises liées à la certitude de transcription (<unclear>), au processus de rédaction du texte source (<add> et <del>), aux citations avec leurs références (<cit>, <quote> et <bibl>) et de termes sélectionnés pour l'indexation (<term>). Ces balises peuvent être utilisées pour créer des plans textuels différents si, par exemple, on ne veut pas que le contenu des citations et des références bibliographiques soit traité de la même façon que les passages rédigés par Flaubert.

La balise <note> nécessite un traitement particulier. Elle sert en effet à marquer deux sortes de segments de texte très différents : les notes originales présentes dans les sources, d'une part, et les commentaires des participants au projet d'édition en ligne d'autre part. Ces derniers portent l'attribut « type » avec la valeur « DBP-footnote ». Puisque le contenu de ces commentaires ne fait pas partie des documents originaux, il convient de les situer « hors-du-texte » et de les exclure de tous les calculs textométriques. Tel est d'ailleurs aussi le cas pour tout ce qui se trouve dans les entêtes des documents <teiHeader>.

Les notes originales peuvent poser un certain problème pour la linéarisation du texte, car elles peuvent se glisser dans les contextes des mots du texte « principal » et ainsi empêcher de calculer la distance correcte entre les occurrences. Dans le cas du corpus DBP, cependant, les notes occupent une place tellement importante dans le corps du texte qu'il nous semble préférable de les « faire remonter » à la surface du texte là où elles se trouvent dans les documents sources.

Un autre cas qui demande une attention particulière lors de l'importation d'un corpus concerne l'usage de différentes langues. La possibilité de savoir pour chaque unité lexicale à quelle langue elle appartient est une condition essentielle pour appliquer les bons modèles linguistiques avec les outils de TAL. La TEI propose une balise spéciale <foreign> pour délimiter des passages en une langue étrangère au sein d'un texte monolingue, mais aussi un attribut global « xml:lang » qui peut être utilisé avec n'importe quelle balise. La langue principale du corpus DBP est le français, mais on y trouve des abréviations latines (*sq.* pour *sequunturque*), des mots en grec et quelques lignes en arabe (notamment sur le folio 221 du volume 4[10]). Malheureusement, ces incrustations étrangères n'ont pas été balisées dans les sources du corpus, ce qui empêche de les traiter correctement lors de l'importation.

En principe, toutes les balises des documents sources contenant du texte peuvent être indexées et exploitées en tant que structures intermédiaires. Cependant, la multiplication de ces structures complexifie la gestion et l'exploitation du corpus. Pour cette raison, il convient de ne garder dans le processus d'importation que les balises qui sont potentiellement utiles pour les requêtes d'exploration. Dans le cas du corpus DBP, on peut notamment éviter d'indexer les balises qui sont très peu utilisées dans l'état actuel des documents sources (par exemple, <closer>, 8 occurrences ; <signed>, 6 occurrences ; <stamp>, 1 occurrence).

Ergonomie du corpus (références et éditions)

La facilité d'exploitation d'un corpus repose en grande partie sur la lisibilité

des pages d'« édition » générées par TXM et sur la précision des références dans les concordances. Dans le cas du corpus DBP, au lieu d'investir dans le « stylage » complexe des pages d'édition TXM, on peut choisir de simplement créer un lien avec le site du projet où pour chaque page s'affichent l'image du manuscrit et les différents types de transcription disponibles.

En ce qui concerne les références des occurrences dans les concordances, on peut se contenter d'afficher l'identifiant de la page correspondante qui intègre en effet l'indication du volume.

Mise en œuvre

Une fois les principes de transposition du corpus source dans le modèle de données TXM définis, il convient de mettre en place la procédure d'importation. À ce jour, le module d'import de sources intitulé « XML/w + CSV » est le seul qui permet d'importer tout type de document XML. C'est celui que nous avons utilisé car aucun des modules plus spécialisés disponibles (tels que « XML-TEI-BFM ») n'était directement compatible avec le schéma de balisage du corpus DBP. Le principal inconvénient du module « XML/w + CSV » est qu'il ne peut pas profiter de la sémantique des balises TEI et demande donc d'effectuer tout un travail de restructuration du corpus en amont du processus d'import.

L'adaptation du corpus à l'analyse par TXM se fait par une transformation XSL paramétrable dans la procédure d'importation. Comme nous l'avons déjà indiqué, une transformation XSL supplémentaire peut être nécessaire pour corriger des erreurs de segmentation lexicale, ce qui a été le cas pour le corpus DBP.

L'insertion automatique dans les pages d'édition TXM des liens vers le site du projet des dossiers de *Bouvard et Pécuchet* a nécessité une opération de recherche et remplacement d'expressions régulières dans les fichiers HTML générés lors de l'importation.

En réalité, l'importation d'un corpus complexe comme DBP se fait rarement d'une seule traite. Des ajustements et des corrections sont souvent nécessaires dans les fichiers sources et dans les traitements automatiques. La plateforme TXM propose par ailleurs de nombreuses fonctionnalités qui facilitent le diagnostic de la qualité du corpus et aident à repérer les erreurs de transcription et d'encodage. Une première exploitation d'un corpus complexe avec la plateforme consiste donc souvent en son diagnostic philologique complet.

Quelques pistes d'exploration

Relecture et vérification

L'exploration d'un corpus sous TXM commence souvent par la fonctionnalité « Description » qui génère une page sommaire indiquant les dimensions du corpus (nombre de mots), les propriétés des unités lexicales (avec un extrait de valeurs), ainsi que toutes les unités de structure avec leurs propriétés, le nombre de valeurs possibles pour chaque propriété et un extrait de valeurs. Dans ce contexte, l'unité textuelle est listée parmi l'ensemble des structures intermédiaires. Les listes de valeurs permettent de détecter facilement d'éventuelles fautes de frappe ou des incohérences au niveau des métadonnées. Par exemple, dans le corpus DBP, on aperçoit un mélange d'anglais et de français dans la liste des types de divisions : on trouve des valeurs « list » et « liste », « text » et « texte ».

Pour vérifier la qualité de la transcription électronique ou de la reconnaissance optique de caractères, on peut recourir à l'outil Index et demander par exemple la liste des formes comprenant à la fois des chiffres et des lettres à l'aide de l'expression CQL (Corpus Query Language)[11] suivante :

```
[word="[a-z]+[0-9].*|[0-9]+[a-z].*"%cd]
```

Les formes les plus fréquentes correspondant à cette requête dans le corpus DBP sont les adjectifs ordinaux (1er, 2e, etc.), mais on trouve également des abréviations de type « 7bre » pour « septembre » et « 9bre » pour « novembre ». Il ne s'agit pas dans ces cas d'erreurs de transcription, mais une régularisation de ces formes pourrait être envisagée pour améliorer les performances de l'annotation linguistique et faciliter les recherches sur le corpus, d'autant plus que la TEI permet très bien de représenter une forme régularisée à côté de l'originale.

Plus bas dans la liste des fréquences, on trouve des formes isolées dues aux erreurs d'OCR comme « iécrimina4ions » ou « cheva1 » (avec le chiffre *1* à la place de la lettre *l*) ou encore des erreurs de tokenisation[12] comme

« 29eANNÉE ».

Deux autres requêtes peuvent aider à repérer des erreurs de tokenisation :

[word=".+-"]

liste toutes les formes contenant un ou plusieurs caractères quelconques suivis d'un trait d'union (de telles formes sont souvent générées lorsque la coupure de mots en fin de ligne, les césures, n'est pas correctement gérée lors de la tokenisation).

[word=".{20,9999}"]

liste toute les formes contenant entre 20 et 9999 caractères, probablement plusieurs mots recollés par erreur. Dans le corpus DBP, on repère ainsi des résolutions d'abréviations considérées comme des mots uniques (*S. A. I.* pour *Son Altesse Impériale*), mais aussi de longues séries de tirets (barres horizontales), qui auraient dû être retirées du plan textuel principal.

Grâce à un mécanisme d'hyperliens, TXM permet de passer en seulement deux double-clics d'une forme de l'index à la page d'édition, ce qui facilite considérablement la localisation des erreurs dans les sources (voir Figures 1, 2 et 3). Ainsi, à partir d'un index de formes finissant par un trait d'union, nous avons affiché une concordance de la forme *ai-*, et puis une page d'édition correspondante, ce qui a permis de repérer une erreur de tokenisation due à une combinaison complexe de balises éditoriales <supplied> et <surplus> avec le saut de ligne <lb/>.

Une fois les erreurs dans les sources corrigées, le corpus peut être réimporté dans TXM pour une exploitation à des fins de recherche.

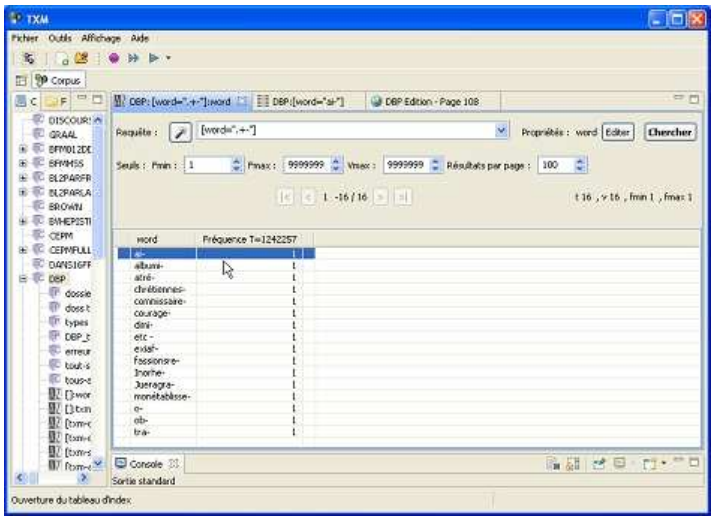


Figure 1. Index des formes contenant au moins un caractère avant le trait d'union final.

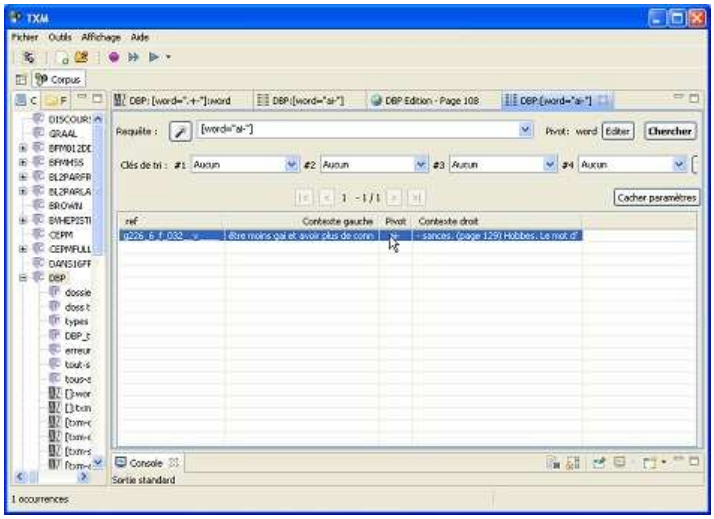


Figure 2. Concordance de la forme *ai-* obtenue par un double-clic sur une ligne de l'index représenté sur la Figure 1.

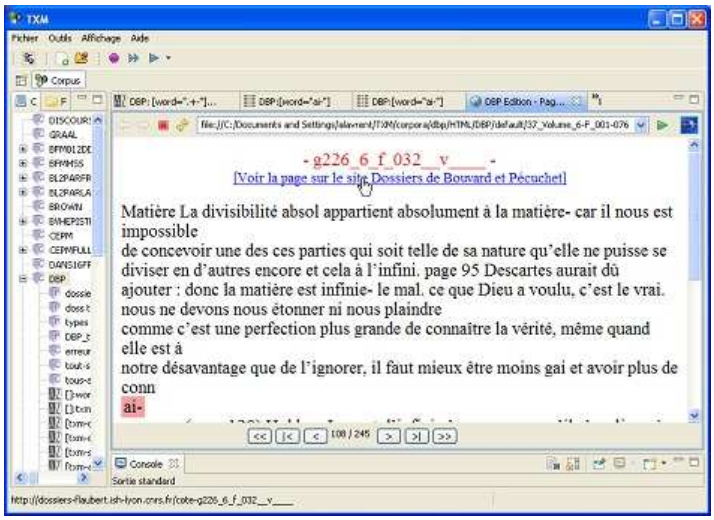


Figure 3. Affichage de l'édition : « retour au texte » par double-clic sur une ligne de la concordance représentée dans la Figure 2. Un hyperlien placé au-dessous du numéro de la page permet d'afficher l'image de la page correspondante sur le site http://www.dossiers-flaubert.fr.

Analyses globales

TXM offre des outils performants pour la création et l'analyse de toutes sortes de contrastes au sein d'un corpus. L'analyse contrastive passe premièrement par la création d'une partition qui divise le corpus selon les critères choisis. Il peut s'agir de métadonnées des unités textuelles, mais aussi de propriétés de structures intermédiaires, voire de requêtes CQL de sélection de mots arbitraires. La façon la plus simple de créer une partition est de choisir une structure et une de ses propriétés dont les valeurs seront à l'origine des parties. TXM crée alors une partition dont chaque partie correspond à une valeur de la propriété. Il convient toutefois de remarquer que, dans l'état actuel du logiciel TXM, la création d'une partition de plus d'une centaine de parties risque d'être longue et son résultat peu exploitable ; il est donc prudent de consulter la liste des valeurs de la propriété choisie (accessible dans le volet « Assisté ») avant de lancer la création d'une partition.

Dans le cas du corpus DBP, nous avons exploré les partitions suivantes :

1. Dossiers thématiques.

Puisque le dossier thématique a été choisi comme « unité textuelle » lors de l'importation du corpus, il s'agit d'une partition de la structure « texte », dont on peut choisir la propriété « id » ou encore la description fournie dans les métadonnées du projet Bouvard. L'un des 52 dossiers thématiques (*Dictionnaire des idées reçues*, cote Ms g 228) n'a pas du tout été transcrit, il n'y a donc que 51 parties contrastables. Deux des 51 dossiers qui restent ont la même description « Pièces diverses », ce qui réduit à 50 le nombre de parties si on base le calcul des partitions sur la description.

La taille des dossiers en nombre de mots transcrits est extrêmement variable : de 13 occurrences-mots dans le dossier « F. 020-059 Manuscrit A » (une section du *Dictionnaire des idées reçues* dont la transcription n'est qu'ébauchée) à plus de 140 000 occurrences dans « Tracts et coupures de journaux ». Un tel déséquilibre peut biaiser certains calculs statistiques, il faut donc procéder avec prudence, sachant que TXM permet de « retailler » les tables lexicales issues des partitions en supprimant ou en fusionnant des colonnes (parties) et des lignes (valeurs de la propriété de mot choisie pour la construction de la table lexicale) et de refaire instantanément les calculs de plans factoriels ou de spécificités.

Une analyse factorielle des correspondances (AFC) basée sur la fréquence des propriétés morphosyntaxiques des mots[13] pour la partition par dossier thématique[14] montre une assez grande homogénéité du corpus (voir Figure 4) : la majorité des dossiers se regroupe dans un « nuage » assez dense près du centre du graphique. On peut tout de même constater que les dossiers qui s'opposent le plus fortement par rapport au premier facteur sont les « Mémoires de Mme Ludovica » d'une part, et les « Références

bibliographiques » et les « Pièces diverses » de l'autre. Si la position particulière de ces deux derniers dossiers n'a rien d'étonnant (on peut s'attendre à y trouver moins de verbes, par exemple), celle, à l'opposé, des « Mémoires de Mme Ludovica » pourrait amener le chercheur à s'intéresser à ce dossier particulier.

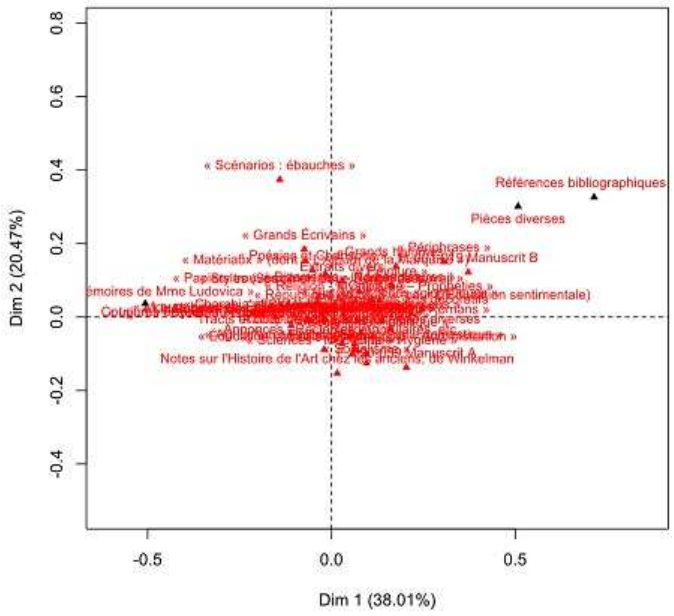


Figure 4. Plan factoriel des dossiers thématiques du corpus DBP représentés par la fréquence des propriétés morphosyntaxiques qu'ils utilisent.

Un autre outil intéressant pour explorer la partition par dossier thématique est le calcul des mots spécifiques à chaque partie[15]. Nous l'avons lancé sur une table lexicale limitée aux lemmes représentés par au moins 50 occurrences dans le corpus. Les résultats se présentent sous la forme d'un tableau qu'on peut trier instantanément par score de spécificité de lemmes dans chaque partie. Sans grande surprise, on trouve en tête de liste les lemmes *Monsieur, Dieu, Religion* dans le dossier « Catholiques », et *enfant, éducation* et *cerveau* dans le dossier « Éducation - Morale - Phrénologie & Administration ». En revanche, il est plus intéressant de trouver *esclave* et *esclavage* dans les lemmes spécifiques du dossier « Religion » et les lemmes *cheval* et *départ* dans le dossier « À classer ». Le premier fait peut caractériser l'attitude de Flaubert vis-à-vis de la religion, tandis que le deuxième peut aider à découvrir la thématique dominante d'un dossier hétérogène.

2. Types de pages

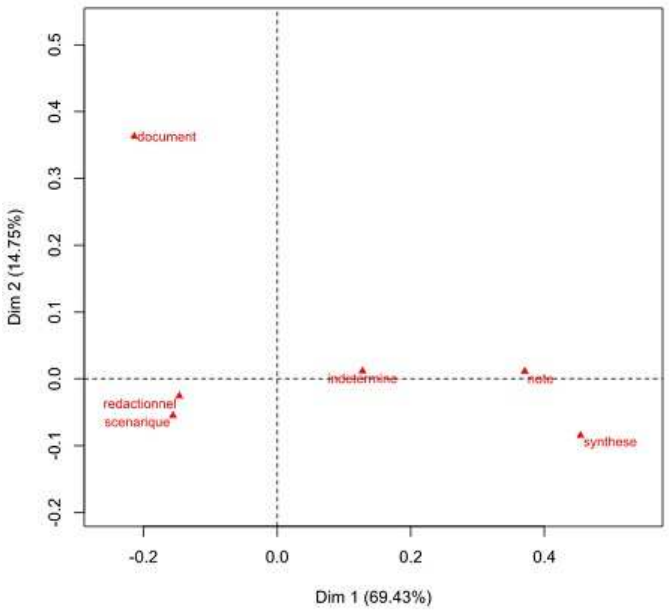
Une autre partition intéressante à analyser dans le corpus DBP est basée sur les types de pages. On constate immédiatement un très grand déséquilibre de parties : près de 650 000 occurrences dans les pages scénariques, 370 000 dans les notes, 88 000 dans les documents et à peine 3 000 occurrences dans les trois types restants (synthèse, indéterminé et rédactionnel). Une AFC (voir Figure 5) sur une table lexicale de lemmes ayant au moins 50 occurrences dans le corpus montre que les pages rédactionnelles, scénariques et les documents se regroupent et s'opposent aux pages de notes et de synthèse selon le premier facteur, tandis que les pages indéterminées occupent une position intermédiaire. Les documents s'opposent à tous les autres types de pages selon le facteur 2. On peut se demander si ne se trouve pas derrière ce phénomène l'opposition entre le style de Flaubert et celui des autres auteurs. Enfin, TXM permet de remplacer l'un des deux premiers facteurs par le facteur 3, qui permet, à son tour, d'opposer les pages de synthèse à tout le reste.

Figure 5. Plan factoriel des types de pages du corpus DBP basé sur les lemmes représentés par au moins 50 occurrences dans le corpus.

3. Scripteurs

Grâce au codage des mains ayant rédigé ou corrigé un passage dans le

corpus DBP, projeté comme « propriété lexicale » sur tous les mots du corpus, il est possible de créer une partition « par main » malgré la diversité des éléments qui portent cette information dans le corpus d'origine. La construction de ce type de partition nécessite le recours à l'interface avancée



de création de partition : chaque partie est définie par une expression CQL de sélection de mots du corpus et porte un nom arbitraire défini par l'utilisateur. L'outil Lexique appliqué à la propriété « dbp-hand » permet tout d'abord d'obtenir la liste des valeurs de cette propriété et leur fréquence. On peut ignorer la valeur « n/a » correspondant aux occurrences dont aucun ancêtre (au sens de l'arborescence XML) ne porte cet attribut. Il s'agit soit de citations imprimées collées sur les pages de dossiers, soit de passages où le balisage des mains n'a pas encore été effectué. Tel est le cas d'une majeure partie du corpus (près de 80 % des occurrences).

Pour le reste du corpus, la plus grande partie est constituée de passages écrits à l'encre par Flaubert (près de 210 000 occurrences), viennent ensuite les passages écrits à l'encre par Edmond Laporte (près de 43 000 occurrences), loin devant l'écriture de Flaubert au crayon (6568 occurrences), les passages dont le scripteur reste indéterminé (6228 occurrences) et les interventions de Jules Duplan (4084 occurrences). Les autres scripteurs sont trop faiblement représentés pour ce genre d'analyse globale (entre 12 et 700 occurrences).

Si on relève les quatre lemmes les plus spécifiques à chacun des scripteurs, on obtient (après filtrage des ponctuations et des mots outils) les indices suivants :

GF-encre : droit (17,1), liberté (15,0), ouvrier (8,7), Dieu (8,5)
GF-crayon : génital (15,9), id. (12,3), Ségur (11,6), contradiction (11,4)
EL-encre : Maistre (63,6), tome (31,8), cité (16,2), Histoire (13,9)
JD : @card@[16] (18,2), me (17,3), Figaro (15,0), je (10,5), ô (9,5)
Indéfini : je (53,3), mon (46,3), me (38,3), vous (13,8)

Ce qui saute aux yeux, c'est la très grande spécificité du pronom de la première personne dans les passages appartenant à un scripteur indéfini. On trouve ce pronom également dans les passages écrits par Jules Duplan, ce qui pourrait éventuellement ouvrir une piste pour l'attribution des passages indéfinis.

On remarque aussi la haute spécificité de l'adjectif *génital* dans les passages écrits au crayon par Flaubert. Enfin, la spécificité des termes *droit* et *liberté* dans les passages écrits par Flaubert à l'encre peut également contribuer à l'analyse littéraire de son œuvre.

Requêtes ciblées

Pour terminer notre exploration textométrique du corpus DBP, nous montrerons quelques exemples de requêtes CQL permettant de retrouver des occurrences de phénomènes précis.

Grâce au traitement des balises <choice> lors de l'importation qui a abouti

à la projection des formes originales irrégulières, erronées ou abrégées vers les propriétés lexicales des formes normalisées, on peut afficher l'index de toutes les corrections en faisant la requête suivante et en sélectionnant « txm-sic » et « word » comme propriétés d'affichage :

```
[txm-sic!="__UNDEF__"]
```

On peut ainsi constater qu'il y a au total 531 occurrences corrigées dans le corpus et que la correction la plus fréquente concerne le remplacement d'un article indéfini ou défini masculin par le féminin (7 occurrences *un* → *une* et 4 occurrences *le* → *la*).

Si on s'intéresse plus particulièrement à la correction de lapsus, il suffit d'ajouter une condition à la requête :

```
[txm-sic!="__UNDEF__" & type="corr-lapsus"]
```

On obtient alors un index de 90 occurrences. Un seul lapsus se reproduit deux fois dans le corpus : *fois* pour *mois*.

Les concordances peuvent servir à approfondir l'analyse des phénomènes repérés lors de l'exploration globale du corpus par des outils quantitatifs. On peut ainsi rechercher toutes les occurrences du lemme *génital*, qui nous a surpris par son haut score de spécificité dans les passages écrits au crayon par Flaubert. En effet, 11 des 25 occurrences de ce terme sont écrites par Flaubert au crayon et se concentrent sur les folios 130-132 du volume 3 (intitulé « Styles (Spécimens de) - Périphrases »). C'est donc une sorte d'indexation thématique réalisée par Flaubert. Les autres occurrences se trouvent dans le volume 7 (« Science – Médecine – Hygiène »), de façon plus espacée. Une seule occurrence se trouve dans le dossier « Éducation – Morale – Phrénologie – Administration », et on peut se demander si cette feuille (fo 169) a été placée dans le bon dossier. Néanmoins, il s'agit d'une page de notes où la citation suivante figure sous le titre « Beauté des hôpitaux » :

> ceux qui sont là (dans les hôpitaux) pour des affections des organes génitaux sont dans la distribution des aliments moins bien traités que les autres (p. 211)

Il ne s'agit donc pas d'une erreur de classement dans ce cas.

L'intérêt de l'exploration textométrique du corpus DBP présenté dans cet article est avant tout d'ordre méthodologique : nous avons pu examiner et partiellement résoudre les difficultés liées à l'importation d'un corpus encodé de façon patrimoniale, dont la relecture et la structuration ne sont pas entièrement terminées, dans un outil d'analyse de corpus textuel. Certains choix de codage du projet Bouvard rendent les tâches de segmentation lexicale et de constitution de plans textuels particulièrement complexes. Par ailleurs, la plateforme TXM dans son état actuel est peu adaptée à la gestion de certaines configurations de corpus (par exemple les corpus composés de plusieurs milliers d'unités textuelles). Malgré toutes ces difficultés, nous avons démontré que l'importation du corpus DBP dans TXM est possible et qu'elle permet d'effectuer un certain nombre d'analyses quantitatives et qualitatives potentiellement intéressantes. Nous ne pouvons qu'espérer que les spécialistes de Flaubert s'approprient l'outil et s'en servent pour répondre à de vraies questions de recherche. Au fur et à mesure de l'amélioration de la qualité de relecture et de structuration du corpus original et de l'évolution de la plateforme TXM, l'exploitation de ce corpus deviendra de plus en plus rapide et pertinente.

NOTES

[1] Gustave Flaubert, *Les dossiers documentaires de Bouvard et Pécuchet*. Édition intégrale balisée en XML-TEI des documents conservés à la bibliothèque municipale de Rouen, accompagnée d'un outil de production de « seconds volumes » possibles, sous la dir. de Stéphanie Dord-Crouslé, 2012, http://www.dossiers-flaubert.fr.

[2] Ces informations sont généralement calculées automatiquement pour chaque mot par les outils de Traitement Automatique de la Langue (TAL) appliqués au corpus (comme les lemmatiseurs automatiques qui associent à chaque mot leur entrée de dictionnaire).

[3] Ces informations sont généralement encodées par les chercheurs lors de l'établissement des textes dans le cadre de la philologie numérique, pratique s'appuyant largement sur la généralisation de l'usage du système de balisage XML qui permet simplement et de façon interopérable à la fois de désigner un

empan textuel donné (une phrase, un paragraphe, une section...) et de lui associer manuellement diverses propriétés.

[4] Voir Jean-Marie Viprey, *Analyses textuelles et hypertextuelles des Fleurs du mal*, Paris, Champion, 2002 ; Étienne Brunet, *Comptes d'auteurs. Études statistiques de Rabelais à Gracq*, Paris, Champion, 2009 ; Bénédicte Pincemin, « Fonctionnalités textométriques pour l'analyse littéraire : possibilités offertes par le logiciel libre TXM », Journée d'étude « Que faire des corpus (une fois) numérisés ? », Caen, 4 juin 2013.

[5] Voir http://txm.sourceforge.net.

[6] Voir http://textometrie.ens-lyon.fr.

[7] Les utilisateurs et les développeurs de la plateforme TXM peuvent dialoguer par le biais de la liste de diffusion de ses utilisateurs :
https://listes.cru.fr/sympa/info/txm-users,
ainsi que de son site wiki communautaire :
https://groupes.renater.fr/wiki/txm-users.

[8] Le dernier état du corpus que nous avons traité date du 23 novembre 2012.

[9] Langage de désignation de balises dans un fichier XML. Voir James Clark, Steve DeRose *et al.*, « XML path language (XPath) version 1.0 », http://www.w3.org/TR/xpath.

[10] Voir http://www.dossiers-flaubert.fr/cote-g226_4_f_221__r____-trud.

[11] Voir la documentation concernant ce langage de requête sur le site de développement de l'environnement Corpus Workbench (http://cwb.sourceforge.net/documentation.php),
ainsi que dans les sections 6 et 7 du *Manuel de TXM* (Serge Heiden *et al.*, *Manuel de TXM*, Version 0.7, Lyon, ENS de Lyon, 2012, http://txm.sourceforge.net/doc/manual/manual.xhtml).

[12] La tokenisation est l'opération de délimitation automatique des formes graphiques des unités lexicales (le découpage des mots).

[13] Article, Déterminant, Substantif, Adjectif, Pronom, Verbe, Adverbe, Préposition... et leurs sous-catégories éventuelles : nombre, genre, temps...

[14] Dans ce calcul, chaque dossier est placé sur le plan factoriel en fonction de coordonnées établies à partir des fréquences d'utilisation de chaque propriété morphosyntaxique (autrement dit, chaque dossier est représenté par l'ensemble des fréquences de propriétés morphosyntaxiques qu'il contient). Deux dossiers ayant des profils de fréquences proches auront des propriétés proches sur le plan factoriel.

[15] L'indice statistique de spécificité d'apparition d'un mot dans une partie du corpus rend compte de l'écart entre la fréquence du mot observée dans la partie et la fréquence théorique du mot à laquelle on pourrait s'attendre dans la partie étant données la fréquence totale du mot dans le corpus, la taille de la partie et la taille du corpus.

[16] Cette notation représente le lemme unique attribué aux adjectifs, déterminants ou pronoms cardinaux par TreeTagger. Nous ne l'avons pas compté parmi les 4 lemmes, mais la spécificité élevé des numéraux dans l'écriture de Jules Duplan peut être intéressante.